Biweekly Milestone Report:
Identifying Context-Based Clusters of Mandarin-English Code-Switching
07-400 Fall 2022

Anna Cai
Mentor: Professor Alan Black, LTI
https://www.andrew.cmu.edu/user/annacai/07400.html

March 4, 2022

1 Major Changes

There have not been any major changes to the goals or implementation of the project.

2 Progress Report

Since the last status update, I have continued testing logistic regression models on the three classification splits (environment, gender, nationality) and determined the most important features for the models' decisions by the magnitude of their weights. Surprisingly, there was not much reliance on POS features, even though logistic regression models trained on only the POS features were able to perform almost as well models trained on all features. This may be due to multiple features encoding similar information about classes. The important features in common with all the splits are fairly generic terms, while some features unique to one split (particularly English terms) are clearly related to particular topics, though the reason why they are considered important by the models is not obvious. I also tested decision trees on the classification problems, but the most important features by Gini importance did not match with those from logistic regression. As the most important features from univariate methods ANOVA feature selection and PCA matched those from logistic regression, it may be possible that the decision tree decisions just do not easily lend themselves to explanation.

Additionally, I looked into using the Chinese-English and English-Chinese dictionaries from MUSE to compare the frequencies between words in different languages with the same meaning. However, the vocabulary in the dictionary did not match many of the words in SEAME, so instead I will likely investigate the frequencies of words with high similarity of multilingual word vectors. As MUSE does not provide already trained word vectors, I may try to use multilingual BERT vectors.

3 Milestone Evaluation

I believe that I have been able to partially meet my milestone goal for this week. While I did not have time to start trying unsupervised clustering techniques yet, I believe I will be able to catch up by the next milestone goal.

4 Surprises

There have not been any major surprises in the project so far.

5 Future Plans

For the next status update, I hope to try topic clustering techniques and investigate the distribution of POS features and classes between clusters. I hope to investigate word frequencies across languages using multilingual word embeddings as well, potentially coming up with features based on vector similarity and difference in frequencies.

6 Milestone Revisions

I will not make any revisions to my milestone goals this time.

7 Resources

So far, I have all the resources I need.